DM manipulation statistique avec R

XinChen22105822

12/10/2021

## 1. Installation des packages

requiredPackages = c('tidyverse', 'languageR', 'Hmisc', 'corrplot', 'broom', 'knitr', 'xtable', 'ggsignif')  
for(p in requiredPackages){  
 if(!require(p,character.only = TRUE)) install.packages(p)  
 library(p,character.only = TRUE)  
}

## 2. Iris {datasets}

data("iris")  
iris %>%  
 View()  
help("iris")

## starting httpd help server ... done

## 3.Description statistiques

Edgar Anderson’s Iris Data est une base de données qui porte sur 50 fleurs de 3 espèces d’iris dont les mesures(en longueur et en largeur) des sépals et des pétales sont données. Le format d’Iris est un dataframe de 150 cellules et de 5 variables(colonnes) “Sepal.Length, Sepal.Width, Petal.Length, Petal.Width, Species”.

Les vaiables

iris %>%   
 group\_by(Species) %>%   
 summary()

## Sepal.Length Sepal.Width Petal.Length Petal.Width   
## Min. :4.300 Min. :2.000 Min. :1.000 Min. :0.100   
## 1st Qu.:5.100 1st Qu.:2.800 1st Qu.:1.600 1st Qu.:0.300   
## Median :5.800 Median :3.000 Median :4.350 Median :1.300   
## Mean :5.843 Mean :3.057 Mean :3.758 Mean :1.199   
## 3rd Qu.:6.400 3rd Qu.:3.300 3rd Qu.:5.100 3rd Qu.:1.800   
## Max. :7.900 Max. :4.400 Max. :6.900 Max. :2.500   
## Species   
## setosa :50   
## versicolor:50   
## virginica :50   
##   
##   
##

Dataframe

str(iris)

## 'data.frame': 150 obs. of 5 variables:  
## $ Sepal.Length: num 5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...  
## $ Sepal.Width : num 3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...  
## $ Petal.Length: num 1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...  
## $ Petal.Width : num 0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...  
## $ Species : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...

## 4.Manipuler la base de données

i.changer la valeur de référence

levels(iris$Species)

## [1] "setosa" "versicolor" "virginica"

iris\_df<- iris %>%  
 mutate(Species = fct\_relevel(Species , "versicolor"))  
levels(iris$Species)

## [1] "setosa" "versicolor" "virginica"

1. selectionner et filtrer les variables

iris %>%  
 select(Sepal.Length, Sepal.Width, Species) %>%  
 filter(Species == "setosa")

## Sepal.Length Sepal.Width Species  
## 1 5.1 3.5 setosa  
## 2 4.9 3.0 setosa  
## 3 4.7 3.2 setosa  
## 4 4.6 3.1 setosa  
## 5 5.0 3.6 setosa  
## 6 5.4 3.9 setosa  
## 7 4.6 3.4 setosa  
## 8 5.0 3.4 setosa  
## 9 4.4 2.9 setosa  
## 10 4.9 3.1 setosa  
## 11 5.4 3.7 setosa  
## 12 4.8 3.4 setosa  
## 13 4.8 3.0 setosa  
## 14 4.3 3.0 setosa  
## 15 5.8 4.0 setosa  
## 16 5.7 4.4 setosa  
## 17 5.4 3.9 setosa  
## 18 5.1 3.5 setosa  
## 19 5.7 3.8 setosa  
## 20 5.1 3.8 setosa  
## 21 5.4 3.4 setosa  
## 22 5.1 3.7 setosa  
## 23 4.6 3.6 setosa  
## 24 5.1 3.3 setosa  
## 25 4.8 3.4 setosa  
## 26 5.0 3.0 setosa  
## 27 5.0 3.4 setosa  
## 28 5.2 3.5 setosa  
## 29 5.2 3.4 setosa  
## 30 4.7 3.2 setosa  
## 31 4.8 3.1 setosa  
## 32 5.4 3.4 setosa  
## 33 5.2 4.1 setosa  
## 34 5.5 4.2 setosa  
## 35 4.9 3.1 setosa  
## 36 5.0 3.2 setosa  
## 37 5.5 3.5 setosa  
## 38 4.9 3.6 setosa  
## 39 4.4 3.0 setosa  
## 40 5.1 3.4 setosa  
## 41 5.0 3.5 setosa  
## 42 4.5 2.3 setosa  
## 43 4.4 3.2 setosa  
## 44 5.0 3.5 setosa  
## 45 5.1 3.8 setosa  
## 46 4.8 3.0 setosa  
## 47 5.1 3.8 setosa  
## 48 4.6 3.2 setosa  
## 49 5.3 3.7 setosa  
## 50 5.0 3.3 setosa

iii.résumés des données (général vs spécifique)

summary(iris)

## Sepal.Length Sepal.Width Petal.Length Petal.Width   
## Min. :4.300 Min. :2.000 Min. :1.000 Min. :0.100   
## 1st Qu.:5.100 1st Qu.:2.800 1st Qu.:1.600 1st Qu.:0.300   
## Median :5.800 Median :3.000 Median :4.350 Median :1.300   
## Mean :5.843 Mean :3.057 Mean :3.758 Mean :1.199   
## 3rd Qu.:6.400 3rd Qu.:3.300 3rd Qu.:5.100 3rd Qu.:1.800   
## Max. :7.900 Max. :4.400 Max. :6.900 Max. :2.500   
## Species   
## setosa :50   
## versicolor:50   
## virginica :50   
##   
##   
##

iris %>%   
 group\_by(Species) %>%  
 summarise(count = n(),  
 range\_Petal.Length = range(Petal.Length),  
 mean\_Petal.Length = mean(Petal.Length),  
 sd\_Petal.Length = sd(Petal.Length),  
 var\_Petal.Length = var(Petal.Length),  
 min\_Petal.Length = min(Petal.Length),  
 max\_Petal.Length = max(Petal.Length),  
 median\_Petal.Length = median(Petal.Length))

## `summarise()` has grouped output by 'Species'. You can override using the `.groups` argument.

## # A tibble: 6 x 9  
## # Groups: Species [3]  
## Species count range\_Petal.Length mean\_Petal.Length sd\_Petal.Length  
## <fct> <int> <dbl> <dbl> <dbl>  
## 1 setosa 50 1 1.46 0.174  
## 2 setosa 50 1.9 1.46 0.174  
## 3 versicolor 50 3 4.26 0.470  
## 4 versicolor 50 5.1 4.26 0.470  
## 5 virginica 50 4.5 5.55 0.552  
## 6 virginica 50 6.9 5.55 0.552  
## # ... with 4 more variables: var\_Petal.Length <dbl>, min\_Petal.Length <dbl>,  
## # max\_Petal.Length <dbl>, median\_Petal.Length <dbl>

## 5.Hypothèse

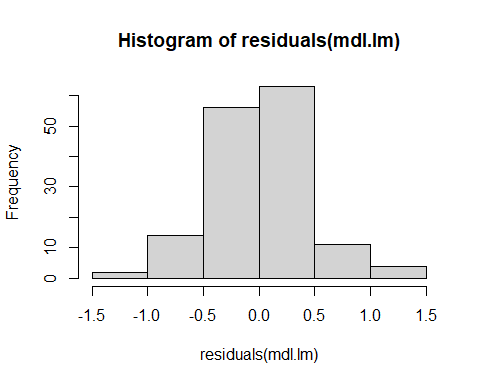
A partir de l’observation sur les données, en particulier le résumé statistique de données par groupes d’espèces, on peut formuler l’hypothèse tel que les variations morphologiques des fleurs d’iris de trois espèces peuvent être quantifié, à savoir que les trois espèces peuvent être discriminé avec des méthodes statistiques. (Si on a les mésures sur les pétales d’une fleur d’iris, on peut prédire leurs espèce avec des méthodes discriminatoires)

i.Premièrement pour tester l’hypothèse, on va construire un modèle linéaire à partir d’un variable dépendant(La longueur de pétales) et le variable indépendant(espèces).

mdl.lm <- iris %>%   
 lm( Petal.Length ~ Species, data = .)  
  
kable(summary(mdl.lm)$coef, digits = 3)

|  | Estimate | Std. Error | t value | Pr(>|t|) |
| --- | --- | --- | --- | --- |
| (Intercept) | 1.462 | 0.061 | 24.023 | 0 |
| Speciesversicolor | 2.798 | 0.086 | 32.510 | 0 |
| Speciesvirginica | 4.090 | 0.086 | 47.521 | 0 |

hist(residuals(mdl.lm))



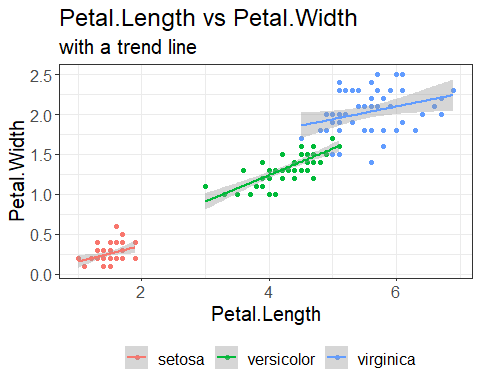
A partir du modèle linéar obtenu, on peut concluer qu’il y a bien une différence importance de la longueur de pétales entre les éspèces. Cependant, est-ce que cette conclusion peut être généralisé pour la largeur de pétales?

En fait, on peut facilement visualiser la réponse à cette question avec les plots linéaires.

## 6.La visualisation

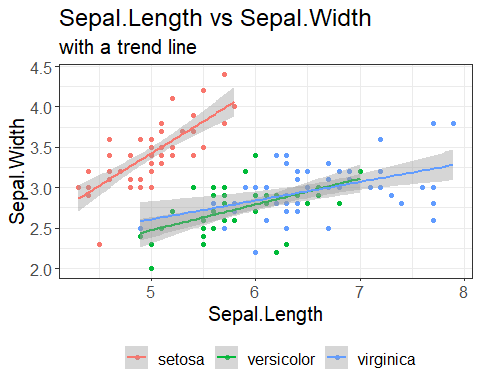
a.Petal.Length vs Petal.Width

## `geom\_smooth()` using formula 'y ~ x'



iris %>%   
 ggplot(aes(x = Sepal.Length,  
 y = Sepal.Width,  
 colour = Species)) +  
 theme\_bw() +  
 geom\_point()+  
 geom\_smooth(method = "lm")+  
 labs(x = "Sepal.Length", y = "Sepal.Width", title = "Sepal.Length vs Sepal.Width", subtitle = "with a trend line") + # add labels  
 theme(text = element\_text(size = 15)) + # increase size of plot  
 theme(legend.position = "bottom", legend.title = element\_blank()) # remove legend title and change position

## `geom\_smooth()` using formula 'y ~ x'



1.Les deux plots(a, b) montre que l’hypothése est prouvé vrai.C’est à dire que si on a les mésures sur les pétales d’une fleur d’iris, on peut prédire leurs espèce avec des méthodes discriminatoires (ses mésures de pétales tombes dans quel groupe d’espèces)

2.à partir de plot a, on voit que la discrimination entre les trois espèces est la plus important lors qu’on prend en compte les mésures de pétals(“Length” et “Width”)

3.à partir de plot b, on voit que la différence entre les espèces est plutôt flou lors qu’on compare les sépales.

## 7. L’Analyse inférentielle

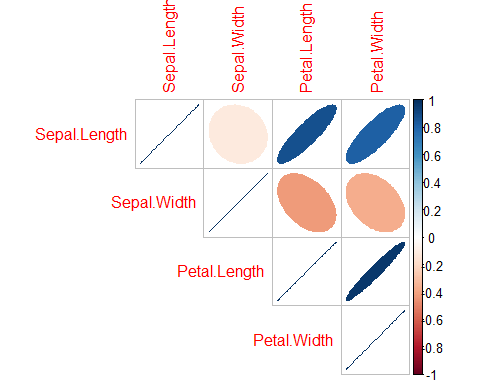
cor(iris$Petal.Length, iris$Petal.Width, method = "pearson")

## [1] 0.9628654

cor.test(iris$Petal.Length, iris$Petal.Width)

##   
## Pearson's product-moment correlation  
##   
## data: iris$Petal.Length and iris$Petal.Width  
## t = 43.387, df = 148, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.9490525 0.9729853  
## sample estimates:  
## cor   
## 0.9628654

corr <-   
 iris %>%   
 select(where(is.numeric)) %>%   
 cor()  
  
corrplot(corr, method = 'ellipse', type = 'upper')



A partir de l’anayse inférentielle, on observe que la corrélation coefficiente est à 0.949 et la valeur p est bien “< 2.2e-16”. C’est à dire qu’il exist bien une corrélation importance entre les deux valeurs(“Petal.Length” et “Petal.Width”).

## 8.Conclusion:

La base de données “iris” a présenté un modèle statistique pour differencier les espèces d’iris. Si on a les mésures sur les pétales(et les sepals?)d’une fleur d’iris, on peut prédire leurs espèce avec des méthodes discriminatoires.